



# A candidate prostate cancer susceptibility gene at chromosome 17p

Sean V. Tavtigian<sup>1</sup>, Jacques Simard<sup>2</sup>, David H.F. Teng<sup>1</sup>, Vicki Abtin<sup>3</sup>, Michelle Baumgard<sup>1</sup>, Audrey Beck<sup>1</sup>, Nicola J. Camp<sup>3,4</sup>, Arlene R. Carillo<sup>1</sup>, Yang Chen<sup>1</sup>, Priya Dayananth<sup>1</sup>, Marc Desrochers<sup>2</sup>, Martine Dumont<sup>2</sup>, James M. Farnham<sup>3</sup>, David Frank<sup>1</sup>, Cheryl Frye<sup>1</sup>, Siavash Ghaffari<sup>1</sup>, Jamila S. Gupte<sup>1</sup>, Rong Hu<sup>1</sup>, Diana Iliev<sup>1</sup>, Teresa Janecki<sup>1</sup>, Edward N. Kort<sup>3</sup>, Kirsten E. Laity<sup>1</sup>, Amber Leavitt<sup>1</sup>, Gilles Leblanc<sup>2</sup>, Jodi McArthur-Morrison<sup>5</sup>, Amy Pederson<sup>1</sup>, Brandon Penn<sup>1</sup>, Kelly T. Peterson<sup>1</sup>, Julia E. Reid<sup>1</sup>, Sam Richards<sup>1</sup>, Marianne Schroeder<sup>1</sup>, Richard Smith<sup>1</sup>, Sarah C. Snyder<sup>1</sup>, Brad Swedlund<sup>1</sup>, Jeff Swensen<sup>1</sup>, Alun Thomas<sup>1</sup>, Martine Tranchant<sup>2</sup>, Ann-Marie Woodland<sup>1</sup>, Fernand Labrie<sup>2</sup>, Mark H. Skolnick<sup>1</sup>, Susan Neuhausen<sup>3</sup>, Johanna Rommens<sup>5</sup> & Lisa A. Cannon-Albright<sup>3,4</sup>

It is difficult to identify genes that predispose to prostate cancer due to late age at diagnosis, presence of phenocopies within high-risk pedigrees and genetic complexity. A genome-wide scan of large, high-risk pedigrees from Utah has provided evidence for linkage to a locus on chromosome 17p. We carried out positional cloning and mutation screening within the refined interval, identifying a gene, *ELAC2*, harboring mutations (including a frameshift and a nonconservative missense change) that segregate with prostate cancer in two pedigrees. In addition, two common missense variants in the gene are associated with the occurrence of prostate cancer. *ELAC2* is a member of an uncharacterized gene family predicted to encode a metal-dependent hydrolase domain that is conserved among eukaryotes, archaebacteria and eubacteria. The gene product bears amino acid sequence similarity to two better understood protein families, namely the PSO2 (SNM1) DNA interstrand crosslink repair proteins and the 73-kD subunit of mRNA 3' end cleavage and polyadenylation specificity factor (CPSF73).

## Introduction

The existence of a familial/genetic component to prostate cancer has long been recognized<sup>1</sup>. Segregation analyses have shown that an autosomal dominant model best fits the pattern of familial clustering of prostate cancer<sup>2</sup>; however, consistently higher risks observed in brothers of prostate cancer affecteds relative to their sons have also led to hypotheses of an X-linked, recessive or imprinted component to the inheritance of prostate cancer susceptibility<sup>3-5</sup>. Consistent with the age-specific risk conferred by the familial colon and breast cancer susceptibility genes, these segregation analyses further predict that the fraction of early onset prostate cancer attributable to deleterious alleles of susceptibility genes is considerably higher than the similarly attributable fraction of late-onset prostate cancer.

So far, localizations of four prostate cancer susceptibility loci, *HPC1* at 1q24 (ref. 6), *PCAP* at 1q42 (ref. 7), *HPCX* at Xq27 (ref. 8) and *CAPB* at 1p36 (ref. 9), have been reported and re-tested in independent data sets. Of these, only the *HPC1* linkage achieved moderately supportive confirmation<sup>10-12</sup>. A recent study presented evidence for linkage to a new locus, *HPC20*, at 20q13 (ref. 5), and a comparison of the 3 complete genome scans published so far<sup>6,13,14</sup>, indicates that there are at least 2 additional regions (11p and 16) that warrant further study. Perhaps the clearest result from these studies is that no single predisposition locus

mapped so far is by itself responsible for a large portion of familial prostate cancer<sup>15</sup>, at least not with the combination of penetrance and clinical distinctiveness that allowed linkage to and positional cloning of the early onset associated colon and breast cancer predisposition genes.

Independent of the linkage analysis and positional cloning approach to identification of prostate cancer susceptibility genes, candidate-gene mutation screening and case-control association study strategies have met with some success. For example, a polymorphic CAG repeat within the androgen receptor (*AR*) ORF, encoding a variable length polyglutamine tract, shows an inverse relationship between repeat length and the transcriptional transactivation activity of the receptor<sup>16</sup>. Accordingly, a series of studies show an association between shorter *AR* CAG repeat length and prostate cancer risk<sup>17,18</sup>. Second, a number of missense variants have been found in the gene encoding steroid 5 $\alpha$ -reductase type II (*SRD5A2*), responsible for conversion of testosterone to the more active androgen dihydrotestosterone in the prostate<sup>19</sup>. One of these variants, Ala49Thr, has been reported to increase the catalytic activity of the enzyme, and is associated with increased risk of advanced prostate cancer<sup>20,21</sup>. Data from these two genes are particularly noteworthy because association with specific variants has been replicated in unrelated population sets and there is biochemical support that the variants would have physiological effects.

<sup>1</sup>Myriad Genetics, Inc., Salt Lake City, Utah, USA. <sup>2</sup>Oncology and Molecular Endocrinology Research Center, Laval University Medical Center (CHUL) and Laval University, Quebec City, Canada. <sup>3</sup>Genetic Epidemiology Group, Department of Medical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, USA. <sup>4</sup>Genetic Research, Intermountain Health Care, LDS Hospital, Salt Lake City, Utah, USA. <sup>5</sup>Program in Genetics and Genomic Biology, Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada. Correspondence should be addressed to S.V.T. (e-mail: [seant@myriad.com](mailto:seant@myriad.com)).

**Table 1 • Two-point linkage evidence at chromosome 17p**

Family resource used to detect linkage to 17p	
Number of pedigrees	33
Total number of cases	338
Total number of typed cases	188
Mean number of cases/pedigree (range)	10.2 (2–29)
Mean number of typed cases/pedigree (range)	5.7 (1–16)
Mean age of typed cases at diagnosis (range)	68.3 (35–88)

Two-point lod scores using Utah age-specific model

Marker	distance (cM)	Max lod* ( $\theta$ )	Heterogeneity lod	( $\alpha$ , $\theta$ )
<i>D17S796</i>	–	0.11 (.37)	0.10	(1.00, 0.4)
<i>D17S952</i>	10.2	0.90 (.17)	0.87	(1.00, 0.2)
<i>D17S786</i>	10.4	0.00 (.50)	0.95	(0.20, 0.0)
<i>D17S945</i>	12.7	0.38 (.28)	1.41	(0.25, 0.0)
<i>D17S520</i>	15.0	0.69 (.26)	0.64	(1.00, 0.3)
<i>D17S974</i>	15.1	1.01 (.20)	1.20	(0.40, 0.01)
<i>D17S1289</i>	15.2	4.53 (.07)	4.43	(1.00, 0.1)
<i>D17S1159</i>	15.4	0.50 (.27)	1.38	(0.25, 0.0)
<i>GATA134G03</i>	15.7	0.48 (.20)	0.78	(0.75, 0.2)
<i>D17S954</i>	16.2	0.00 (.50)	0.11	(0.40, 0.2)
<i>D17S969</i>	18.2	0.54 (.21)	0.55	(0.85, 0.2)
<i>D17S799</i>	22.0	0.30 (.26)	0.44	(0.70, 0.2)
<i>D17S921</i>	25.2	1.41 (.10)	1.42	(0.95, 0.1)
<i>D17S953</i>	29.2	1.04 (.25)	0.94	(1.00, 0.3)
<i>D17S925</i>	31.2	0.02 (.45)	0.00	(1.00, 0.0)
<i>D17S798</i>	36.2	0.02 (.43)	0.02	(1.00, 0.4)

\*Maximum lod scores interpolated using the standard quadratic function.

It is difficult to estimate the relative contribution that low-frequency, high-risk variants (analogous to mutations in *MLH1* and *MSH2* or *BRCA1* and *BRCA2*) versus higher-frequency, moderate-risk sequence variants (such as short alleles of the *AR* CAG repeat and the Thr49 variant of *SRD5A2*) make to the population risk of prostate cancer. If moderate-risk sequence variants are responsible for a significant fraction of the attributable risk of prostate cancer, then models of the genetic component of familial prostate cancer may need to incorporate both linkage evidence at high-risk susceptibility loci and consideration of contributing moderate-risk sequence variants.

The Utah population provides a unique resource for examining the genetic basis of disease. Founded by approximately 10,000 immigrants of British, Scandinavian and German origin<sup>22</sup> who were followed by more than 50,000 migrants from the same areas in the next generation, the population demonstrates typical Northern European gene frequencies<sup>23</sup> and normal levels of inbreeding<sup>24</sup>. The population structure allows ascertainment of extended high-risk pedigrees containing many cases as units instead of by expansion from individual probands<sup>11,22</sup>. These pedigrees are an extremely powerful resource for linkage studies, and they also allow analysis of segregation of moderate-risk sequence variants through multiple generations of cases and their unaffected relatives.

## Results

### Linkage analysis

We originally carried out a genome-wide search for prostate cancer predisposition loci using a small set of high-risk prostate cancer pedigrees from Utah and a set of 300 polymorphic markers. The pedigrees were not selected by age of onset, but were a subset of families ascertained using the Utah Population Database<sup>11,22</sup>. The first 8 pedigrees analyzed gave suggestive evidence of linkage on chromosome 17p near marker *D17S520*, although significance was not established. We then increased the density of markers in the region and expanded the analysis to 33 pedigrees (Table 1). Using a dominant model integrated with Utah age-specific incidence, the analysis yielded a maximum 2-point lod score of 4.5 at marker *D17S1289*,  $\theta=0.07$ , and a maximum 3-point lod

score of 4.3 using the markers *D17S1289* and *D17S921* (Table 1). The linkage evidence was not due to one or a few families with high lod scores, but rather was distributed across the family set. Based on these data, we initiated a positional cloning project, focusing on the interval between *D17S1289* and *D17S921*.

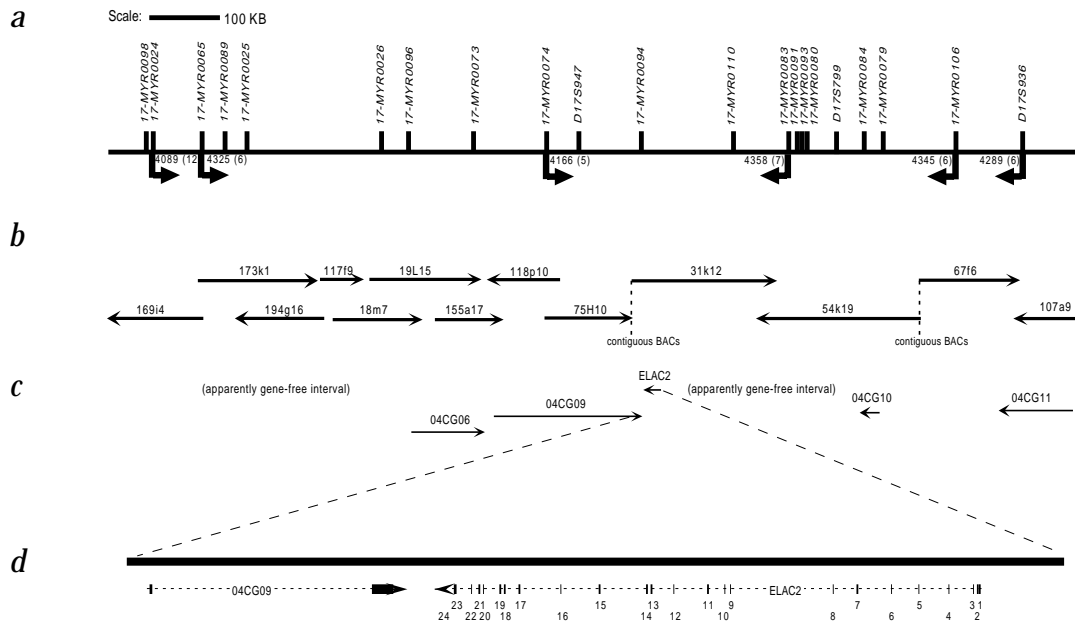
To refine the localization of the implied susceptibility gene, we added an additional 94 families, making a total of 127 (Table 2); these were typed for markers at both this and the *HPC1* locus. The overall data set neither provides significant lod score evidence for linkage on chromosome 17p nor provides sufficient evidence for *de novo* identification of the *HPC1* locus<sup>11</sup>. For instance, the maximum 17p multipoint lod score<sup>25</sup> for the additional 94 families was 0.44 at the marker *17-MYR0110*.

Early in our analysis, we observed that, at both 17p and *HPC1*, many of our pedigrees segregate haplotypes that are shared by four or more cases, but also contain enough non-carrying cases with respect to either locus to eliminate any linkage evidence within the pedigree, as estimated by lod score. For instance, 12 affected individuals from kindred 4333 share a *HPC1* haplotype and 9 affected in kindred 4344 share a 17p haplotype, but neither pedigree shows lod score evidence for linkage at either locus. Although we recognize that this phenomenon may simply be due to lack of linkage, we hypothesized that the underlying cause is actually genetic complexity that is greater than the linkage models can accommodate. We subsequently used multipoint haplotyping software<sup>25</sup> to define segregating haplotypes, and then classified those haplotypes into three groups, depending on strength of evidence: group 1 haplotypes, used for both localization and mutation screening, were defined as haplotypes shared by 4 or more cases and giving a lod score greater than or equal to 1.0 in the pedigree in which they were identified, or haplotypes shared by 6 or more cases irrespective of lod score; group 2 haplotypes, used for mutation screening only, were defined as haplotypes shared by 4 cases with  $0.5 < \text{lod score} < 1.0$  in the pedigree in which they were identified, or haplotypes shared by 5 cases with lod score less than 1.0; and, finally, haplotypes that failed to meet any of the above criteria.

Considering group 1 and 2 haplotypes together, supporting evidence for *HPC1* and 17p are comparable: 43 haplotypes at *HPC1* versus 42 at 17p, and 258 affected haplotype carriers at *HPC1* versus 232 at 17p. Focusing on the group 1 haplotypes, evidence at *HPC1* is relatively stronger: 26 group 1 haplotypes at *HPC1* versus 18 at 17p, with an average of 7.2 affected carriers per *HPC1* group 1 haplotype versus 6.6 per 17p group 1 haplotype. But there is one other critical difference between the linkage evidence for the two regions. Meiotic recombinant mapping using the *HPC1* group 1 haplotypes has so far failed to define a consistent region. This result is also evident in the ICPCG *HPC1* study<sup>12</sup>, in which most of the evidence for linkage comes from a combination of the Utah and Hopkins data sets, but the locations with the best evidence for linkage in each of the individual sets map approximately 15 cM apart. In contrast, recombinant mapping in affected carriers of 17p group 1 haplotypes defined a consistent region (Fig. 1a). As a result, we were able to focus our contig assembly, transcript map development and mutation screening efforts on an approximately 1.5-Mb interval flanked by *17-MYR0024* and *D17S936* (Fig. 1b,c).

**Table 2 • Family resource genotyped for the association tests**

Number of pedigrees	127
Total number of cases	2,402
Total number of typed cases	700
Total number of typed pedigree unaffecteds	3,295
Total number of typed divergent controls	243
Mean number of cases/pedigree (range)	18.3 (3–74)
Mean number of typed cases/pedigree (range)	5.5 (1–34)
Mean age of typed cases at diagnosis (range)	66.5 (39–88)



**Fig. 1** Recombinant, physical and transcript map centered at the human *ELAC2* locus on chromosome 17p. **a**, Genetic markers and recombinants. (Microsatellite markers developed at Myriad are given as 17-MYR####.) Nested within the arrows that represent meiotic recombinants is the number of the kindred in which the recombinant occurred and, in parentheses, the number of cases carrying the haplotype on which the recombinant occurred. **b**, BAC contig tiling path across this interval. The T7 end of each BAC is denoted with an arrowhead. **c**, Transcription units identified in the interval. **d**, Expanded view of a 40-kb segment at the SP6 end of BAC 31k12 showing the relative positions of 2 exons of the gene 04CG09 and all of the coding exons of *ELAC2*.

© 2001 Nature Publishing Group <http://genetics.nature.com>

One of the genes mapping within this interval encodes a protein sharing amino acid sequence similarity with members of the NCBI Cluster of Orthologous Groups<sup>26</sup> COG1234, typified by the uncharacterized *Escherichia coli* ORF ElaC and the *Saccharomyces cerevisiae* ORF YKR079C. On mutation screening this candidate gene from the genomic DNA of prostate cancer cases carrying 17p group 1 haplotypes, we found a germline insertion mutation, 1641insG, in a carrier from kindred 4102. Following detection of this insertion and the predicted frameshift, the gene, referred to as *ELAC2*, was subjected to sequence and genetic analyses.

### Sequence analysis

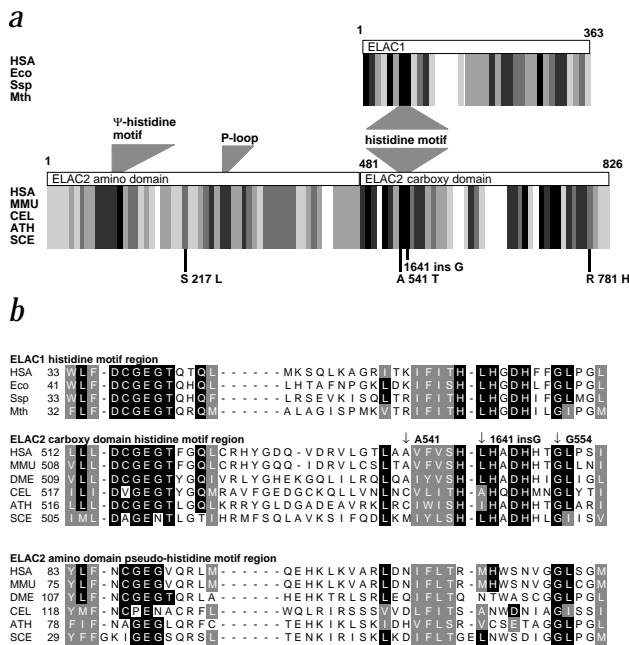
We assembled the human and mouse *ELAC2* cDNA sequences from a combination of ESTs, hybrid selected clones and 5'-RACE products. Conceptual translation of the human cDNA sequence yielded a protein of 826 amino acids; parsing the cDNA sequence across the corresponding genomic sequence revealed 24 coding exons (Fig. 1d). Mouse *Elac2* encodes a protein of 831 residues in 25 exons. BLAST (ref. 27) searches of the *ELAC2* sequence against GenBank revealed a single ortholog in *S. cerevisiae* (YKR079C), a single ortholog in *Caenorhabditis elegans* (CE16965, CELE04A4.4), and a single ortholog in *Drosophila melanogaster* (juvenile hormone-inducible protein-1; ref. 28). Two related sequences were found in both *Schizosaccharomyces pombe* and *Arabidopsis thaliana*. Alignment of representative orthologs revealed good conservation near the amino termini and a series of high similarity segments across the carboxy-terminal half of the proteins (Fig. 2a), although cross-kingdom sequence identity between the complete ORFs is only in the range of 18% to 25%.

Hybridization of RNA blots to labeled fragments of human *ELAC2* cDNA revealed a single transcript of approximately 3 kb (Fig. 3), in agreement with our assembled cDNA sequence of 2,970 bp. The transcript was detected in all tissues surveyed and, like *BRCA1* and *BRCA2* mRNAs, was most abundant in testis. There was no evidence from RNA blots, EST sequences or RT-PCR experiments of alternative splicing of the transcript.

In surveying dbEST, we identified a small number of human and rabbit ESTs originating from a second, related gene. The human cDNA sequence of this related gene, encoding a predicted polypeptide of 363 residues, was also assembled from a combination of ESTs and 5'-RACE products. Radiation hybrid mapping placed the gene at approximately 365 cR on chromosome 18. When this sequence along with representative sequences from a eubacterium (*E. coli* ElaC), a cyanobacterium (*Synechocystis* sp. gj2500943/SLR0050) and an archaeobacterium (*M. thermoautotrophicum* gj2622965) was added into the multiprotein alignment (Fig. 2a), it became apparent that two distinct groups of proteins were represented: a group of larger proteins (800–900 aa) restricted to the eukaryotes; and a group of smaller proteins (300–400 aa) that align with the C-terminal half of the former group and include sequences from the eukaryotes, eubacteria and archaeobacteria. As the 363-residue human protein falls into this second group and is more similar to *E. coli* ElaC than is *ELAC2*, we will refer to it as *ELAC1*.

The alignment revealed that the human and mouse *ELAC2* proteins share a potential ATP/GTP-binding site motif A (P-loop; ref. 29) that begins at residue 276 (Fig. 2a) in the human sequence. A semi-conservative substitution occurs in the P-loop of the *S. cerevisiae* ortholog, whereas more divergence is evident in the *C. elegans* and *A. thaliana* sequences (Fig. A, see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)). On this basis, it is not clear whether the mammalian *ELAC2* proteins possess a nucleotide binding site, or share a spurious motif match.

The alignment also revealed a histidine motif,  $\phi\phi\phi[S/T]HxHxDHxxG$  (where  $\phi$  can be any large hydrophobic residue), near the N terminus of the *ELAC1* group, and in the middle portion of the *ELAC2* group. This motif is similar to the histidine motif found in the metallo- $\beta$ -lactamases<sup>30</sup> and indicates, in accord with the annotation for COG1234 (<http://www.ncbi.nlm.nih.gov/COG/index.html>), that the proteins are metal-dependent hydrolases. We also observed that the sequences within which the histidine motif is embedded show alignment with *ELAC2* N-terminal sequences (Fig. 2b), leading us to predict



**Fig. 2** Multiple protein alignment of ELAC1/2 family members. **a**, The graphic is a representation of the multiple sequence alignment presented in Fig. A (see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)). That alignment has been transformed to grayscale in windows of ten amino acids wherein darker shades represent higher, and lighter shades lower, levels of sequence similarity. ELAC1 orthologs were selected from human (HSA), *E. coli* (Eco), the blue-green algae *Synechocystis* sp. (Ssp), and the archaeobacterium *Methanobacterium thermoautotrophicum* (Mth). ELAC2 orthologs were selected from human (HSA), mouse (MMU), *C. elegans* (CEL), *A. thaliana* (ATH) and *S. cerevisiae* (SCE). Positions of the pseudo-histidine motif, the P-loop, and the histidine motif are indicated by gray-shaded pointers. **b**, Multiple protein sequence alignment demonstrating similarity between the histidine motif and surrounding region in ELAC1 orthologs, the histidine motif and surrounding region in ELAC2 orthologs, and the N domain pseudo-histidine motif and surrounding region in ELAC2 orthologs. In addition to the sequences used in (a), the sequence of the ELAC2 ortholog from *D. melanogaster* (DME) was also included. Alignments were based on BLASTp searches and then optimized by inspection. The positions of Ala541, 1641insG and Gly554 in human ELAC2 are marked by ↓.

that some structural feature of the ELAC2 proteins is repeated. Even so, the histidine motif itself is not conserved within these N-terminal sequences; consequently, the N-terminal domain would not necessarily retain metal-dependent hydrolase activity.

In addition to the histidine motif and the local sequence context in which it is embedded, BLAST searches of GenBank, combined with iterative motif searches<sup>31</sup> using the eMOTIF SCAN website (<http://dna.stanford.edu/scan>), revealed two other conserved homologous groups that share additional sequence features with ELAC1/2 (Fig. B, see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)). One such group encodes the PSO2 (or SNM1) DNA inter-strand crosslink repair proteins<sup>32,33</sup>, present only in eukaryotes. The second group encodes the 73-kD subunit of the mRNA cleavage and polyadenylation specificity factor<sup>34–36</sup> (CPSF73); members of the latter group are present in both eukaryotes and archaeobacteria, as well as a cyanobacterium. The ELAC1/2 family and the PSO2 and CPSF73 groups are equally similar to each other (Fig. B, see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)), and they were originally placed in a single COG (ref. 31). Notably, all three groups have three or four conserved histidine or cysteine residues, past the histidine motif, that lie within these shared regions and can be aligned across the gene family (Fig. B, see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)). The arrangement is reminiscent of the binuclear zinc-binding active site of some metallo-β-lactamases<sup>37,38</sup> and the shared similarity between the metallo-β-lactamases and glyoxalase II (ref. 30). These sequence similarities lead to three predictions. First, extended similarity between the ELAC1/2, PSO2 and CPSF73 proteins indicates that they share a domain of approximately 300 residues constituting a metal-dependent hydrolase that coordinates 2 divalent cations in its active site. Second, the overall fold of this domain is likely to be similar to that of the metallo-β-lactamases. Third, similarity between the N and C portions of the ELAC2 proteins indicates that they are com-

posed of two structurally similar domains and arose from a direct repeat/duplication of an ancestral ELAC1-like gene.

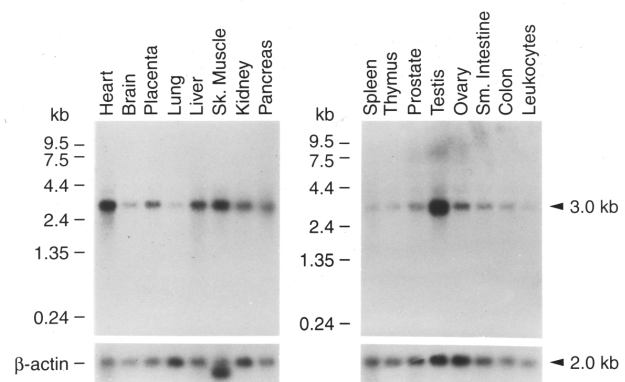
Although the *S. cerevisiae* CPSF73 ortholog YSH1 (BRR5) is an essential gene, PSO2 is not. Given that the ELAC1/2 family is phylogenetically conserved and that *S. cerevisiae* encodes only a single member, YKR079C, we determined whether YKR079C is essential. We performed a one-step gene disruption of YKR079C using URA3 as a selectable marker in yeast diploid cells. Two heterozygous mutant strains were sporulated for tetrad dissection. Each tetrad yielded one or two viable haploid colonies; these were all URA<sup>-</sup> and YKR079C wild type. Thus, we concluded that YKR079C, like YSH1, is an essential gene.

A number of members of the ELAC1/2 family are annotated in GenBank as sulfatases or sulfatase homologs. The annotation appears to be assigned through sequence similarity to the *atsA* gene of *Alteromonas carrageenovora*, which has been demonstrated to have aryl sulfatase activity *in vitro*<sup>39</sup>. The *AtsA* protein contains a histidine motif, and the *E. coli* protein most similar to *AtsA* is *ElacC*; therefore, *A. carrageenovora AtsA* is most likely a member of the ELAC1 group (BLASTp and alignment not shown). Accordingly, ELAC1/2 family members should be tested for aryl sulfatase activity, even though *AtsA* does not contain any of the typical sulfatase motifs listed by PROSITE.

### Mutation screening

Kindred 4102 was ascertained as a high-risk cluster with 8 prostate cancer cases in a 3-generation pedigree. Genotyping revealed that 6 of the 8 cases shared a chromosome 17p haplotype (this haplotype is crosshatched in Fig. 4a). Mutation screening from lymphocyte DNA of the youngest (age 46 at diagnosis) affected carrier of this shared haplotype, 4102.58 (that is, kindred 4102, individual 58; Fig. 4a), revealed an insertion/frameshift, 1641insG, in *ELAC2*. Insertion of a nucleotide between residues 1,641 and 1,642 shifts

**Fig. 3** Analysis of *ELAC2* expression in human tissues. Top, multiple-tissue northern filters probed with the human *ELAC2* ORF. Note that a 3-kb *ELAC2* transcript is detected in all tissues. Bottom, the same filters probed with *ACTB* (β-actin) as a loading control.

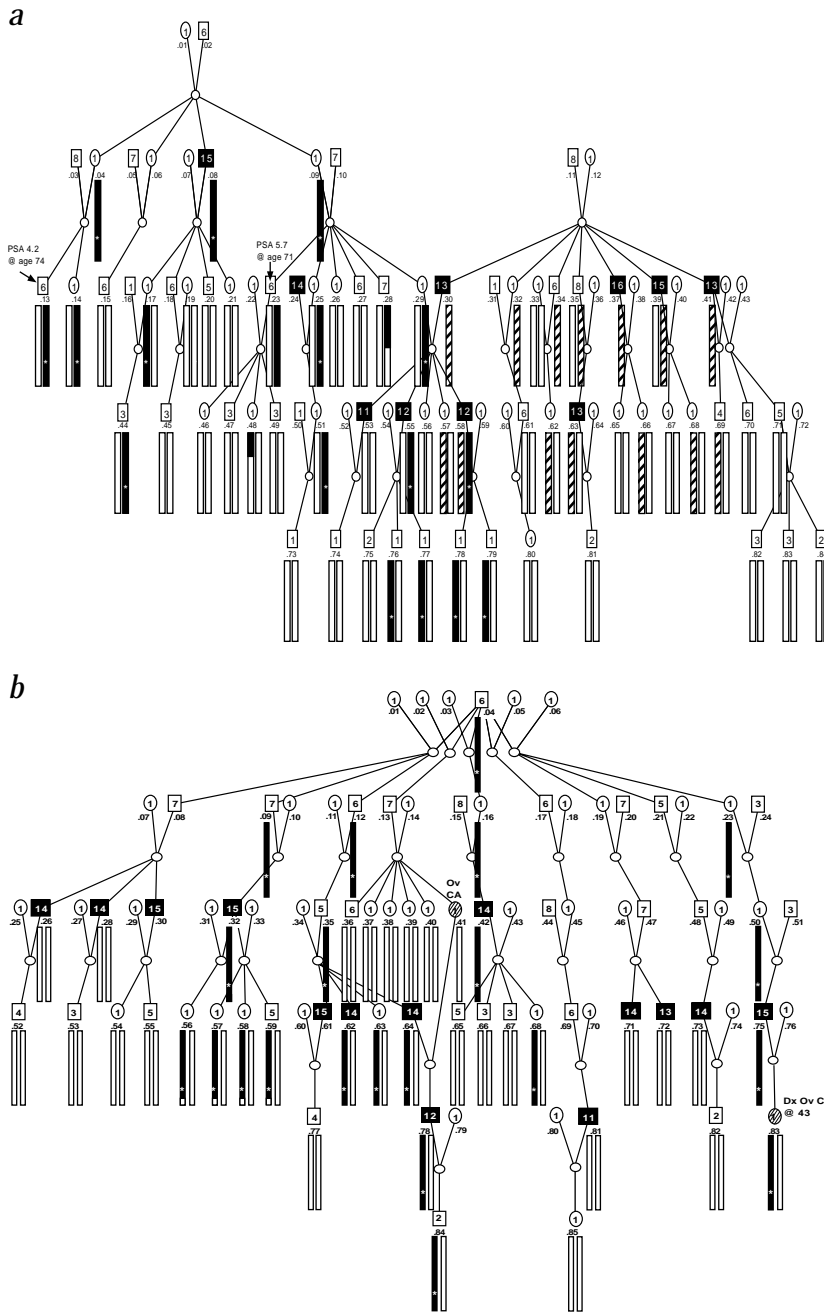


the reading frame of *ELAC2* after Leu547, leading to termination after the miscoding of 67 residues. Because the frameshift occurs within the histidine motif and eliminates over one-third of the polypeptide, including several strongly conserved segments, the kindred 4102 mutation is predicted to be disruptive to the protein. A test for segregation revealed that the mutation was not on the shared 17p haplotype inherited from the father, but rather was inherited through the patient's mother, 4102.29. Expansion of the pedigree revealed that her maternal uncle, 4102.08, was an obligate frameshift carrier. He was diagnosed with and died of prostate cancer at age 76. In all, there are 5 male frameshift carriers over age 45

in the pedigree. Of these, 3 have prostate cancer, the fourth has a total prostate specific antigen (PSA) of 5.7 at age 71, and the fifth has a PSA of 4.2 at age 74 (Fig. 4a).

As the mutation 1641insG was found in an individual with early onset prostate cancer, we screened an additional 45 prostate cancer cases with early age at diagnosis (Dx≤55 years), irrespective of evidence of linkage to any locus, for mutations in *ELAC2*. We identified an alteration, Arg781His, in individual 4289.78, who was diagnosed with prostate cancer at age 50. On expansion of his pedigree (Fig. 4b), the mutation was traced back four generations to 4289.04, who had affected descendants from five known wives.

Prostate cancer cases carrying the missense change have been found among the descendants from three of these five marriages. Of 14 prostate cancer cases in the pedigree, 6 carry the missense change, 6 are non-carriers and the remaining 2 are unknown. In addition, a female carrier of this missense change, 4289.83, was diagnosed with ovarian cancer at age 43. Within the generations with phenotype information, there are only 2 unaffected males over age 45 who carry the mutation: 4289.59 (PSA of 0.6 at age 60) and 4289.35, who died of a heart attack at age 62. We have no additional information on 4289.35, but two sons and a grandson are carriers diagnosed with prostate cancer. Expansion from a single affected mutation carrier to a pedigree with a lod score of 1.3 provides supportive evidence that the mutation is deleterious. The missense change occurs in a very highly charged stretch of amino acid residues near the C terminus of the protein. Arg781 is conserved in mouse, and the charge character of the sequence segment is conserved in *C. elegans* (Fig. 2a; Fig. A, see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)).

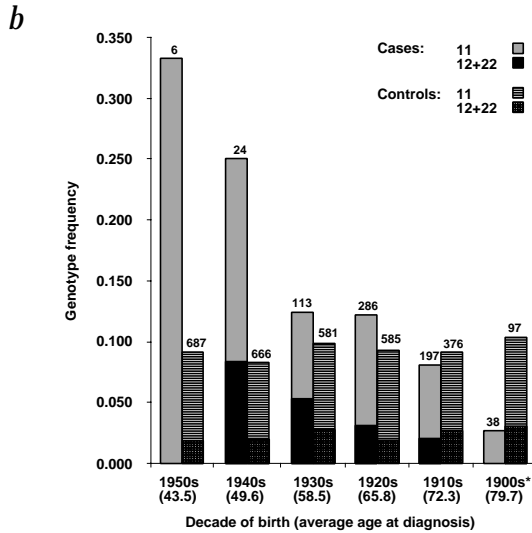


**Fig. 4** Kindreds 4102 and 4289. The pedigrees have been genotyped over a 20-cM interval extending from *D17S786* to *D17S805*. Haplotypes are represented by the bars; the black haplotype segregating in each pedigree is the mutation-bearing chromosome. The relative position of *ELAC2* is denoted by an asterisk. **a**, Kindred 4102. The black bar denotes the 1641insG bearing haplotype. Individuals .28 and .48 carry part of the frameshift haplotype, but neither carry the frameshift due to recombination events. There are no data to distinguish which of the founders, individuals .01 and .02, carried the frameshift. The second shared haplotype in kindred 4102 is denoted by a crosshatched bar. Again, there are no data to distinguish which of the founders, individuals .11 and .12, carried the haplotype. **b**, Kindred 4289. The black bar denotes the His781 bearing haplotype. Individuals .56, .57, .58 and .59 share a recombinant chromosome that carries the His781 missense change. **c**, Liability class table for the pedigree drawings of (a,b).



**a**

Allele	Defining sequence variant(s)	Note
0	Ser217 + Ala541	matches mouse at polymorphic positions
1	Leu217 + Ala 541	allele frequency = 26.0%
2	Leu217 + Thr 541	allele frequency = 3.9%



Identification of two mutations segregating with disease constitutes provocative evidence that *ELAC2* is a prostate cancer susceptibility gene. After screening 42 haplotypes with evidence for linkage at 17p, however, we have found only these 2 high-risk mutations. Thus, it seems that only a small fraction of prostate cancer pedigrees segregate obvious mutations in the *ELAC2* coding sequence. We do not yet know what fraction of the pedigrees harbor gene rearrangements or regulatory mutations.

#### Common missense changes

When our original set of linked pedigrees was screened for mutations in *ELAC2*, we observed several occurrences of the non-conservative missense change Ser217Leu. Like the common human allele, the mouse and *C. elegans* residues at this position are also serine. Although the sequence of the segment around Ser217 is not well conserved, its hydrophilic character is (Fig. 2a; Fig. A, see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)); thus substitution of Ser217 with a bulky hydrophobic residue may result in structural consequences to the protein.

We analyzed this sequence variant in our pedigree cases, unaffected pedigree members and an unrelated set of males who have no diagnosis of cancer (divergent controls). The total number of individuals typed exceeded 4,000 (Table 2), with an overall allele frequency of 30% for Leu217. A logistic regression was performed for disease status to delineate effects of genotypes versus birth year (a demographic datum collected on all participants). We observed a significant interaction between genotype and birth year ( $P=0.027$ ). Genotype frequencies differed across birth cohorts for cases, but appeared more uniform in the unaffected controls (Fig. 5b). There are a number of possible explanations for this observation, including any difference in cases and controls born before 1920 that would result in a difference in the likelihood of being genotyped, resulting in an ascertainment bias. In any event, it is well known that main effects, in the presence of interactions, are uninterpretable and stratification based on the interacting variable is in order. We therefore chose to analyze the effect of genotype in

Fig. 5 *ELAC2* allele definitions and genotype frequencies. **a**, Allele definitions based on the common missense variants Ser217Leu and Ala541Thr. One would predict a fourth allele, Ser217/Thr541. However, we observed very strong disequilibrium between Thr541 and Leu217, and this observation has been repeated in an independent study<sup>10</sup>. Thus the frequency of a Ser217/Thr541 allele must be very low. **b**, Recessive genotype frequencies by birth cohort. The frequency of Leu217 homozygotes, subdivided into Thr541 non-carriers and carriers, is given for cases and controls in decade birth cohorts. The number atop each column indicates the number of individuals in that birth cohort. The birth cohort marked 1900s\* also includes 1 case and 3 controls born in the 1890s.

individuals born after 1919, as the data suggest that a different risk pattern may exist for individuals born before this date.

We were able to make two comparisons: cases versus divergent controls and cases versus unaffected pedigree members. In these analyses (Table 3), prostate cancer patients born between 1920 and 1959 were more frequently Leu217 homozygotes than either the divergent controls (57/429 versus 9/148,  $P$  value=0.026) or the unaffected pedigree members (57/429 versus 220/2371,  $P$  value=0.013). Thus the association tests are consistent with a hypothesis that the Leu217 variant is either deleterious or in disequilibrium with another deleterious variant.

Upon mutation screening *ELAC2* in the set of early onset prostate cancer cases, we also observed several instances of a second non-conservative missense change, Ala541Thr. Ala541 lies at the amino border of the histidine motif (Fig. 2) at a position that may have bearing on the predicted enzymatic activity of the protein; in *S. cerevisiae* PSO2, a missense change at Gly256, a conserved residue corresponding to Gly554 at the carboxy edge of the *ELAC2* histidine motif (Fig. 2; Fig. B, see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)), confers a temperature-sensitive phenotype<sup>33</sup>. The Thr541 variant has been examined in the same set of cases, unaffected pedigree members and divergent controls, in which it has an overall allele frequency of 4%. Thr541 is in strong disequilibrium with Leu217; in fact, we have yet to observe a chromosome that carries Thr541 that does not also carry Leu217. Another logistic regression was performed to investigate effects of genotypes at Ala541Thr. Again, a significant interaction between genotype and birth year was found ( $P$  value=0.003), along with evidence for an effect of genotype at Ala541Thr on disease status.

The frequency of Thr541 is significantly higher in prostate cancer cases than in divergent controls, such that the variant appears to be dominant and deleterious (carrier frequency of 42/429 versus 5/148,  $P$  value=0.022; Table 3). In the comparison between cases and unaffected pedigree members, an effect at Thr 541 only emerged when Leu217 homozygotes were subdivided into Thr541 carriers and non-carriers. In this comparison, the presence of Thr541 on a homozygous Leu217 background is associated with a higher odds ratio than simple homozygosity at Leu217 (2.0 versus 1.4), and the model remains statistically significant ( $P$  value=0.017, trend test  $P$  value=0.004; Table 3). We interpret these data to indicate that the Leu217/Thr541 allele is more deleterious than the allele bearing just Leu217. Given the strong disequilibrium between Thr541 and Leu217, it is possible that the increased risk we observe for Leu217 homozygotes is entirely due to linkage disequilibrium with Thr541. But subset analyses in both the case versus divergent control and case versus pedigree unaffected comparisons show significantly increased risk for Leu217 homozygotes irrespective of Thr541 genotype (data not shown).

One of the two Thr541 homozygotes that we observed among our prostate cancer cases was 4289.78, the first observed carrier of the His781 variant. We established by segregation analysis that the His781 variant occurs on a chromosome that also carries Leu217 and Thr541; therefore, the allele of *ELAC2* shared by at least six prostate cancer cases and one ovarian cancer case within



**Table 3 • Tests for association between *ELAC2* genotypes and prostate cancer**

1920–1959 birth cohort: cases versus divergent controls					
Test 1: Leu217 recessive					
Genotype groups	cases	controls			
(00, 01, 02)	372	139			
(11, 12, 22)	57 (13.3%)	9 (6.1%)	Odds ratio	2.4	
			P value	0.0261	
Test 2: Thr541 dominant					
(00, 01, 11)	387	143			
(02, 12, 22)	42 (9.8%)	5 (3.4%)	Odds ratio	3.1	
			P value	0.0223	
Tests 1 and 2 combined					
(00, 01)	347	137			
(02, 11, 12, 22)	82 (19.1%)	11 (7.4%)	Odds ratio	2.9	
			P value	0.0014	
1920–1959 birth cohort: cases versus pedigree unaffecteds					
Test 1: Leu217 recessive					
(00, 01, 02)	372	2,151			
(11, 12, 22)	57 (13.3%)	220 (9.3%)	Odds ratio	1.5	
			P value	0.0135	
Trend test					
(00, 01, 02)	372	2,151			
(11)	40 (9.3%)	170 (7.2%)	Odds ratio 11	1.4	
(12, 22)	17 (4.0%)	50 (2.1%)	Odds ratio 12,22	2.0	
			P value	0.0173	
			trend statistic	8.09	
			P value	0.004	

The divergent controls are a set of male spouses from cancer pedigrees, but not prostate cancer pedigrees, who themselves have no diagnosis of cancer. Allele definitions are given in Fig. 5, and the genotypes used in these contingency tables are based on those allele definitions.

kindred 4289 actually bears all three missense changes. Across our set of 127 extended pedigrees, we have found 6 examples, including kindred 4289, of Leu217/Thr541 alleles shared IBD by 4 or more prostate cancer cases within a single pedigree. Two of these haplotypes include six cases, and none include seven or more cases. Compared with the five other instances of four or more cases sharing a Leu217/Thr541 allele IBD, the kindred 4289 allele appears, by itself, the most penetrant. Although we have insufficient data to statistically exclude the possibility that His781 is simply a neutral variant in linkage disequilibrium with Leu217 and Thr541, we suspect that His781 contributes to the phenotype conferred by the triple missense allele.

### Discussion

Observation of segregating mutations in *ELAC2* in kindreds 4102 and 4289, plus association between two common missense changes, Leu217 and Thr541, with a diagnosis of prostate cancer, constitutes strong evidence that certain variants in this gene confer increased risk of prostate cancer. In addition, the association observed between the Leu217/Thr541 allele and a diagnosis of prostate cancer has recently been replicated in an independent set of unrelated cases and controls<sup>40</sup>. Therefore, we conclude that *ELAC2* is a prostate cancer susceptibility gene.

We have found four germline variants that alter the amino acid sequence encoded by *ELAC2*; one frameshift and three missense changes. The 164insG frameshift found in kindred 4102 will clearly disrupt protein function, but this is not so obviously true of the three missense changes we have detected. Results from our association tests and segregation analyses, however, are consistent with an additive hypothesis to explain the relative strengths of the missense-bearing alleles that we have observed. Substitution of Leu for Ser217 may change the character of a normally hydrophilic segment of the protein; the phenotype conferred is

sufficiently modest that it is only clearly detected when the variant (in the absence of Thr541 and/or His781) is present in the homozygous state. Ala541 is immediately adjacent to the histidine motif, is conserved in mouse and *Drosophila*, and is the most common residue at this position in the *ELAC2*, *CPSF73* and *PSO2* orthologous groups (Fig. 2; Fig. B, see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)). Although threonine is observed at this position in other genes containing histidine motifs, it is rare or absent at this position in these three closely related groups of genes. Thus, from sequence conservation considerations, it is reasonable that the Leu217/Thr541 allele should be more deleterious than Leu217 alone, apparently sufficiently deleterious to be detected in association tests using a dominant model<sup>40</sup>. The kindred 4289 allele carries all three missense changes, Leu217, Thr541 and His781. Examination of the pedigree indicates that the allele is dominant and segregates with prostate cancer, consistent with conferring higher risk than the commonly observed Leu217/Thr541 allele. The youngest affected carrier of the triple missense allele, 4289.78, is homozygous for Leu217 and Thr541. Thus his mother, the second ovarian cancer case in the pedigree, is an obligate carrier of a Leu217/Thr541 allele. The observation of two ovarian cancer cases in this pedigree, both of whom carry deleterious alleles of *ELAC2*, is consistent with the possibility that the phenotype conferred by deleterious variants in this gene is not restricted to prostate cancer susceptibility.

In addition to increased prostate cancer risk for carriers, we observed a birth-year by genotype interaction among both Leu217 homozygotes and Thr541 carriers. Although this interaction may turn out to be an artifact of ascertainment, it may also turn out to be bona fide. Sequence analysis suggests that *ELAC2* encodes a metal-dependent hydrolase. Therefore, one could well imagine differential interaction of variants of this gene product with some environmental exposure not prevalent during the early 1900s.

The contributions of the *AR* CAG repeat and the *SRD5A2* Ala49Thr missense change to prostate cancer risk were first detected in association studies using sporadic cases and unaffected controls. However, straightforward deduction from sib-pair analyses would predict that such sequence variants should be enriched among affected sibs versus isolated cases, and it follows that such sequence variants should contribute to a larger fraction of familial than truly sporadic prostate cancer cases. Thus one might expect specific genotypes of moderate risk susceptibility genes such as *AR*, *SRD5A2* and the common missense changes in *ELAC2* to confound linkage studies aimed at detecting and localizing less prevalent, high-risk susceptibility genes. Inclusion of genotype information from pedigree members at multiple moderate risk loci may allow refined definition of the liability classes used by multipoint linkage software, thereby increasing the power of the analysis. Stratification of cases by genotype would also facilitate positional cloning projects by providing another criterion by which to distinguish between true recombinant carriers and confounding sporadic cases.

The genetic data presented here demonstrate that there are deleterious sequence variants in *ELAC2* that contribute to prostate cancer risk. Elucidating the functional alteration by which moderate risk sequence variants such as Leu217 or Thr541 contribute to a late-onset pathology could prove difficult because manifestation may be quite subtle. But a frameshift leading to protein truncation within the likely active site of an enzyme should have a more easily detected effect on cell physiology. Conservation of the C-terminal domain of the gene through the eubacteria and archaeobacteria, combined with the observation that *S. cerevisiae* YKR079C is essential, emphasize that the functions of the *ELAC1* and *ELAC2* gene families are of fundamental biological interest.





## Methods

**Linkage analysis.** All participants signed informed consent documents. This research project has the approval of the University of Utah School of Medicine Institutional Review Board. We confirmed 97% of cancer cases through medical records (and/or through the Utah Cancer Registry for prostate cancer cases diagnosed in Utah). Two-point linkage analysis was performed with the package LINKAGE (ref. 41) using the FASTLINK implementation<sup>42,43</sup>. The statistical analysis for the inheritance of susceptibility to prostate cancer used a model that assumes age-specific incidence rates from the Utah Cancer Registry, and a relative risk of 2.5 for first-degree relatives. Susceptibility to prostate cancer was assumed due to a dominant allele with a population frequency of 0.003. The details of the model have been thoroughly defined<sup>11</sup>. Marker allele frequencies were estimated from unrelated individuals present in the pedigrees. Linkage in the presence of heterogeneity was assessed by the admixture test<sup>44</sup> (A-test) using HOMOG, which postulates two family types, linked and unlinked. Within the 17p interval examined in detail, marker positions were estimated using CRIMAP (ref. 45) and three-point linkage analysis was performed using VITESSE (ref. 46).

**Physical mapping.** We purified BAC DNA on columns (Qiagen) and sequenced ends with dye terminator chemistry (ABI) on ABI 377 sequencers. DNA sequences at the SP6 and T7 ends of isolated BAC clones were used to develop STSs that were used for mapping and contig extension. Greater than 95% sequence coverage of the BAC tiling path was obtained by sequencing plasmid sublibraries generated from these clones. The sequence data obtained were assembled into contigs using Acembly, version 4.3 (U. Sauvage, D. Thierry-Mieg and J. Thierry-Mieg). Subsequently, a complete sequence of this interval was released by the MIT genome center.

**cDNA assembly.** We identified ESTs cognate to the human *ELAC2* locus by BLASTn (ref. 27) analysis of genomic sequences from BAC 31k12 against GenBank. ESTs cognate to mouse *Elac2* were identified in the same search. cDNA sequences of these two genes were extended into their respective 5' UTRs by biotin capture 5'-RACE (ref. 47) from preparations of human prostate and mouse embryo cDNA. Human-gene-specific reverse primers were (primary amplification) 5'-(biotin)-TGACGC CTTCTCCACAGT-3' and (secondary amplification) 5'-(phosphate)-GTACCCGCTGCCACCAC-3'. Mouse-gene-specific reverse primers were 5'-(biotin)-CAGAACACATTTGGGAAGC-3' (primary amplification) and 5'-(phosphate)-GATGTTGTCCAAGCGAGC-3' (secondary amplification). Assignment of the translation start codon in human and mouse was confirmed by the identification of a 5' UTR in-frame stop codon that is conserved between the two species.

**Northern blots.** We purchased multiple tissue northern (MTN) filters (Clontech), which are loaded with 2 µg per lane of poly(A)<sup>+</sup> RNA derived from a number of human tissues. <sup>32</sup>P-random-primer labeled probes corresponding to a 2.5-kb cDNA fragment of the coding human *ELAC2* sequence (exon 1 to exon 24) and *ACTB* (β-actin) were used to probe the same set of filters successively. Prehybridization and hybridization were performed with ExpressHyb (Clontech) as recommended by the manufacturer. The membranes were washed twice in 2×SSC/0.1% SDS at 20 °C for 15 min followed by 2 stringency washes in 0.1×SSC/0.1% SDS at 50 °C for 20 min.

**Multiple protein sequence alignments.** For the alignment of Fig. 2b, shading criteria were ≥80% identity (white on black) or conservative substitution (white on gray) for all ELAC1 and ELAC2 histidine motif region sequences. Shaded positions in the ELAC1 and ELAC2 histidine motif sequences were propagated into the ELAC2 amino domain histidine motif-like sequences. For the alignment of Fig. A (see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)), shading criteria were identity (white on black) or conservative substitution (white on gray) for all ELAC2 sequences with a residue at that position, with four of the five sequences actually having to have a residue at that position. Shaded positions in the ELAC2 sequences were propagated into the ELAC1 sequences. For the alignment of Fig. B (see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)), shading criteria were identity or conservative substitution across two of the three (CPSF73, PSO2, ELAC2) protein families represented.

***S. cerevisiae* gene mutation.** *URA3* was PCR amplified with tailed primers resulting in a product flanked by 42 bp of YKR079C coding sequences (aa 3–16 and 818–831). The resulting PCR product was trans-

formed into yeast diploid strain YPH501 (Stratagene); *URA*<sup>+</sup> clones were screened for disruption by the presence of a shorter PCR product at the YKR079C locus. The mutant clones were further confirmed by sequencing the shorter PCR product for the presence of *URA3* sequences. Two heterozygous mutant strains were sporulated and tetrads dissected. Each tetrad yielded 1 or 2 viable colonies. These were genotyped at YKR079C and tested for growth on *URA*<sup>-</sup> plates.

**Mutation screening.** Using genomic DNAs from prostate kindred members as templates, we carried out nested PCR amplifications to generate PCR products to screen for mutations in *ELAC2*. Sequences of a set of primers sufficient to mutation screen the entire ORF from genomic DNA are given (Table A, see [http://genetics.nature.com/supplementary\\_info/](http://genetics.nature.com/supplementary_info/)). Using the outer primer pair for each amplicon (1A-1P, that is, forward A and reverse P of amplicon 1), Genomic DNA (10–20 ng) was subjected to a 25-cycle primary amplification, after which the PCR products were diluted 45-fold and reamplified using nested M13-tailed primers (1B-1Q, 1C-1R, that is, nested forward B and nested reverse Q of amplicon 1 or nested forward C and nested reverse R of amplicon 1) for another 23 cycles. In general, samples were amplified with *Taq* Platinum (Life Technologies) DNA polymerase; cycling parameters included an initial denaturation step at 95 °C for 3 min, followed by cycles of denaturation at 96 °C (12 s), annealing at 55 °C (15 s) and extension at 72 °C (30–60 s). After the PCR reactions, excess primers and deoxynucleotide triphosphates were digested with exonuclease I (United States Biochemicals) and shrimp alkaline phosphatase (Amersham). PCR products were sequenced with M13 forward and reverse fluorescent (Big Dye, ABI) dye-labeled primers on ABI 377 sequencers. We obtained more than 95% double-strand sequence coverage for the entire ORF of all samples screened.

**Association tests.** STSs for Ser217 Leu and Ala541Thr were amplified by allele-specific PCR using fluorescently labeled oligonucleotides<sup>48</sup>. Allele calls were made with an automated genotyping system (Myriad). Genotype calls required good allele calls at both markers. Logistic regression analyses were performed using the SPSS statistical software package. The  $\chi^2$  statistics for the 2×2 contingency tables were calculated with the Yates correction. The trend statistic for the 3×2 contingency table was calculated with the Cochran-Armitage trend test<sup>49,50</sup> using a simple linear trend (0,1,2) for the row scores.

**GenBank accession numbers.** Human *ELAC1* cDNA, AF308695; mouse *Elac1* cDNA, AF308697; human *ELAC2* cDNA, AF304370; mouse *Elac2* cDNA, AF308696; chimpanzee *ELAC2*, cDNA, AF308698; gorilla *ELAC2* cDNA, AF308694; human *ELAC2* genomic sequence, AC005277.

## Acknowledgments

We thank M. Anderson, M. Boyack, S. Faught, C. Hansen, M. Higbee (deceased), M. Jost, R. Nelson, K. Nguyen, T. Peterson, L. Steele, T. Tran and A. Zeller for technical assistance; F. Bazan for structural modeling; A. Bansal for suggestions with regards to the association studies; K. Heichman, J. Shaw and B. Bleazard for assistance with the yeast knockout; and D. Shattuck for a critical reading of the manuscript. This work was supported by grants CA62154 and CA64477 from the National Institutes of Health, funding from Schering Plough Research Institute, and by Endorecherche. The research was also supported by the Utah Cancer Registry, which is funded by contract no. N01-PC-67000 from the National Cancer Institute with additional support from the Utah Department of Health and the University of Utah.

Received 8 May; accepted 22 December 2000.

1. Woolf, C.M. An investigation of the familial aspects of carcinoma of the prostate. *Cancer* **13**, 361–364 (1960).
2. Schaid, D.J., McDonnell, S.K., Blute, M.L. & Thibodeau, S.N. Evidence for autosomal dominant inheritance of prostate cancer. *Am. J. Hum. Genet.* **62**, 1425–1438 (1998).
3. Narod, S.A. *et al.* The impact of family history on early detection of prostate cancer. *Nature Med.* **1**, 99–101 (1995).
4. Monroe, K.R. *et al.* Evidence of an X-linked or recessive genetic component to prostate cancer risk. *Nature Med.* **1**, 827–829 (1995).
5. Berry, R. *et al.* Evidence for a prostate cancer-susceptibility locus on chromosome 20. *Am. J. Hum. Genet.* **67**, 82–91 (2000).
6. Smith, J.R. *et al.* Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. *Science* **274**, 1371–1374 (1996).
7. Berthon, P. *et al.* Predisposing gene for early-onset prostate cancer, localized on chromosome 1q42.2–43. *Am. J. Hum. Genet.* **62**, 1416–1424 (1998).
8. Xu, J. *et al.* Evidence for a prostate cancer susceptibility locus on the X



- chromosome. *Nature Genet.* **20**, 175–179 (1998).
9. Gibbs, M. *et al.* Evidence for a rare prostate cancer-susceptibility locus at chromosome 1p36. *Am. J. Hum. Genet.* **64**, 776–787 (1999).
  10. Cooney, K.A. *et al.* Prostate cancer susceptibility locus on chromosome 1q: a confirmatory study. *J. Natl. Cancer Inst.* **89**, 955–959 (1997).
  11. Neuhausen, S.L. *et al.* Prostate cancer susceptibility locus HPC1 in Utah high-risk pedigrees. *Hum. Mol. Genet.* **8**, 2437–2442 (1999).
  12. Xu, J. & ICGPG Combined analysis of hereditary prostate cancer linkage to 1q24–25: results from 772 hereditary prostate cancer families from the International Consortium for Prostate Cancer Genetics. *Am. J. Hum. Genet.* **66**, 945–957 (2000).
  13. Suarez, B.K. *et al.* A genome screen of multiplex sibships with prostate cancer. *Am. J. Hum. Genet.* **66**, 933–944 (2000).
  14. Gibbs, M. *et al.* A genomic scan of families with prostate cancer identifies multiple regions of interest. *Am. J. Hum. Genet.* **67**, 100–109 (2000).
  15. Ostrander, E.A. & Stanford, J.L. Genetics of prostate cancer: too many loci, too few genes. *Am. J. Hum. Genet.* **67**, 1367–1375 (2000).
  16. Chamberlain, N.L., Driver, E.D. & Miesfeld, R.L. The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function. *Nucleic Acids Res.* **22**, 3181–3186 (1994).
  17. Giovannucci, E. *et al.* The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. *Proc. Natl. Acad. Sci. USA* **94**, 3320–3323 (1997).
  18. Stanford, J.L. *et al.* Polymorphic repeats in the androgen receptor gene: molecular markers of prostate cancer risk. *Cancer Res.* **57**, 1194–1198 (1997).
  19. Makridakis, N. *et al.* A prevalent missense substitution that modulates activity of prostatic steroid 5 $\alpha$ -reductase. *Cancer Res.* **57**, 1020–1022 (1997).
  20. Makridakis, N.M. *et al.* Association of mis-sense substitution in SRD5A2 gene with prostate cancer in African-American and Hispanic men in Los Angeles, USA. *Lancet* **354**, 975–978 (1999).
  21. Jaffe, J.M. *et al.* Association of SRD5A2 genotype and pathological characteristics of prostate tumors. *Cancer Res.* **60**, 1626–1630 (2000).
  22. Skolnick, M.H. The Utah genealogical data base: a resource for genetic epidemiology. in *Banbury Report No. 4: Cancer Incidence in Defined Populations* (eds. Cairns, J., Lyon, J.L. & Skolnick, M.H.) 285–297 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1980).
  23. McLellan, T., Jorde, L.B. & Skolnick, M.H. Genetic distances between the Utah Mormons and related populations. *Am. J. Hum. Genet.* **36**, 836–857 (1984).
  24. Jorde, L.B. & Skolnick, M.H. Demographic and genetic application of computerized record linking: the Utah Mormon genealogy. *Information Sciences Humaines* **56-57**, 105–117 (1981).
  25. Thomas, A., Gutin, A., Abkevich, V. & Bansal, A. Multipoint linkage analysis by blocked Gibbs sampling. *Stat. Comp.* **10**, 259–269 (2000).
  26. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
  27. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
  28. Dubrovsky, E.B., Dubrovskaya, V.A., Bilderback, A.L. & Berger, E.M. The isolation of two juvenile hormone-inducible genes in *Drosophila melanogaster*. *Dev. Biol.* **224**, 486–495 (2000).
  29. Walker, J.E., Saraste, M., Runswick, M.J. & Gay, N.J. Distantly related sequences in the  $\alpha$ - and  $\beta$ -subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**, 945–951 (1982).
  30. Melino, S., Capo, C., Dragani, B., Aceto, A. & Petruzzelli, R. A zinc-binding motif conserved in glyoxalase II,  $\beta$ -lactamase, and arylsulfatases. *Trends Biol. Sci.* **23**, 381–382 (1998).
  31. Nevill-Manning, C.G., Wu, T.D. & Brutlag, D.L. Highly specific protein sequence motifs for genome analysis. *Proc. Natl. Acad. Sci. USA* **95**, 5865–5871 (1998).
  32. Haase, E., Riehl, D., Mack, M. & Brendel, M. Molecular cloning of SNM1, a yeast gene responsible for a specific step in the repair of cross-linked DNA. *Mol. Gen. Genet.* **218**, 64–71 (1989).
  33. Niegemann, E. & Brendel, M. A single amino acid change in SNM1-encoded protein leads to thermoconditional deficiency for DNA cross-link repair in *Saccharomyces cerevisiae*. *Mutat. Res.* **315**, 275–279 (1994).
  34. Chanfreau, G., Noble, S.M. & Guthrie, C. Essential yeast protein with unexpected similarity to subunits of mammalian cleavage and polyadenylation specificity factor (CPSF). *Science* **274**, 1511–1514 (1996).
  35. Jenny, A., Hauri, H.P. & Keller, W. Characterization of cleavage and polyadenylation specificity factor and cloning of its 100-kilodalton subunit. *Mol. Cell. Biol.* **14**, 8183–8190 (1994).
  36. Jenny, A., Minvielle-Sebastia, L., Preker, P.J. & Keller, W. Sequence similarity between the 73-kilodalton protein of mammalian CPSF and a subunit of yeast polyadenylation factor I. *Science* **274**, 1514–1517 (1996).
  37. Carfi, A. *et al.* X-ray structure of the ZnII  $\beta$ -lactamase from *Bacteroides fragilis* in an orthorhombic crystal form. *Acta Crystallogr. D. Biol. Crystallogr.* **54**, 45–57 (1998).
  38. Fabiane, S.M. *et al.* Crystal structure of the zinc-dependent  $\beta$ -lactamase from *Bacillus cereus* at 1.9 Å resolution: binuclear active site with features of a mononuclear enzyme. *Biochemistry* **37**, 12404–12411 (1998).
  39. Barbeyron, T., Potin, P., Richard, C., Collin, O. & Kloareg, B. Arylsulphatase from *Alteromonas carrageenovora*. *Microbiology* **141**, 2897–2904 (1995).
  40. Rebbeck, T.R. *et al.* Association of HPC2/ELAC2 genotypes and prostate cancer. *Am. J. Hum. Genet.* **67**, 1014–1019 (2000).
  41. Lathrop, G.M., Lalouel, J.M., Julier, C. & Ott, J. Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci. USA* **81**, 3443–3446 (1984).
  42. Cottingham, R.W., Jr., Idury, R.M. & Schaffer, A.A. Faster sequential genetic linkage computations. *Am. J. Hum. Genet.* **53**, 252–263 (1993).
  43. Schaffer, A.A., Gupta, S.K., Shriram, K. & Cottingham, R.W., Jr. Avoiding recomputation in linkage analysis. *Hum. Hered.* **44**, 225–237 (1994).
  44. Ott, J. Linkage probability and its approximate confidence interval under possible heterogeneity. *Genet. Epidemiol. Suppl.* **1**, 251–257 (1986).
  45. Lander, E.S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* **84**, 2363–2367 (1987).
  46. O'Connell, J.R. & Weeks, D.E. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genet.* **11**, 402–408 (1995).
  47. Tavtigian, S.V. *et al.* The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nature Genet.* **12**, 333–337 (1996).
  48. Rust, S., Funke, H. & Assmann, G. Mutagenically separated PCR (MS-PCR): a highly specific one step procedure for easy mutation detection. *Nucleic Acids Res.* **21**, 3623–3629 (1993).
  49. Cochran, W.G. Some methods of strengthening the common chi-squared tests. *Biometrics* **10**, 417–451 (1954).
  50. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386 (1955).